Vers une Ingénierie des Connaissances Personnelles – Étude de cas pour l'organisation des collections musicales

Nicolas Greffard, Pascale Kuntz, Éric Languénou

Vers une Ingénierie des Connaissances Personnelles – Étude de cas pour l'organisation des collections musicales

Nicolas Greffard, Pascale Kuntz, Éric Languénou

LINA/UNIVERSITÉ DE NANTES
2, rue de la Houssinière, BP 92208
44322 Nantes Cedex 03, France
[nicolas.greffard, pascale.kuntz,eric.languenou]@univ-nantes.fr

Résumé : Le traitement de nos informations numériques personnelles est devenu une tâche majeure de nos vies contemporaines et la complexité des informations stockées nécessite des approches permettant de structurer et de gérer les connaissances associées à ces dernières. Dans cette communication, nous nous focalisons sur une étape amont préalable à la proposition d'outils d'assistance personnalisés : l'identification des processus d'organisation (ici de classement) dans un environnement familier. Le terrain d'expérimentation est celui des collections musicales personnelles stockées sur les disques durs de nos ordinateurs. Une analyse de la structuration de ces collections nous a permis de mettre en évidence un ensemble restreint de stratégies d'organisation qui affinent celles étudiées dans la littérature pour d'autres types de collections. Notre discussion ouvre sur les nouveaux défis associés à l'émergence du syndrome du « big data » à l'échelle individuelle.

Mots-clés : données personnelles, classification, collections musicales.

1 Introduction

Les terrains privilégiés par l'ingénierie des connaissances (IC) sont les entreprises et les administrations (Charlet, 2005). Ce positionnement s'explique à la fois par l'historique des systèmes experts et des systèmes à base de connaissances mais aussi par les besoins actuels des institutions qui ne cessent de croître. Pourtant la définition de l'IC n'est pas si restrictive dans sa cible et peut s'appliquer aujourd'hui, nous semble-t-il, au champ individuel. En effet, le traitement de nos informations numériques personnelles est devenu une tâche majeure de nos vies contemporaines et la volumétrie et la complexité des informations stockées nécessitent des méthodes et techniques permettant de structurer et de gérer les nouvelles connaissances associées à ces informations ; c'est-à-dire une ingénierie des connaissances adaptée à l'échelle personnelle. Si nous n'avons pas trouvé trace de ces préoccupations dans les actes de la conférence IC de cette dernière décennie, des questionnements sur l'impact du virage numérique dans le traitement des connaissances individuelles sont au cœur des débats de la communauté structurée autour du « Personal Information Management » (PIM) (Jones, 2008) dont l'attention a porté initialement principalement sur les documents, les courriels et les pages Web. L'impact du numérique sur l'accès et le stockage dans la sphère privée stimule aujourd'hui les recherches sur l'organisation des collections personnelles (Lee, 2011) associées notamment aux bibliothèques numériques, aux photothèques et vidéothèques personnelles. Dans cet article, nous nous penchons sur un champ qui a été très peu étudié : celui de l'organisation des collections musicales personnelles. Cette absence relative contraste avec l'importance de la musique dans nos vies quotidiennes (DeNora, 2000). Aujourd'hui l'attention médiatique porte majoritairement sur le

streaming car ses revenus sont en croissance mais, pourtant, le téléchargement représente encore 49% du marché et une enquête récente de l'Hadopi confirme l'attachement individuel des jeunes à la constitution des bibliothèques numériques de biens culturels. Dans ce contexte, nos travaux se focalisent ici sur l'organisation des discothèques numériques personnelles.

Plus précisément, l'objectif de cet article est d'identifier les modes d'organisation mis en œuvre dans le classement des collections musicales stockées sur nos disques personnels. Cette identification des comportements développés dans un environnement familier nous semble un préalable nécessaire à la proposition d'outils d'assistance personnalisés. Après un état de l'art synthétique sur les travaux – assez rares – consacrés à ces collections, nous présentons une étude effectuée sur un échantillon de disques durs recueillis essentiellement chez des jeunes adultes et nous discutons des perspectives nouvelles que cette démarche ouvre en ingéniérie des connaissances.

2 État de l'art

De nombreux outils de gestion des données musicales ont été développés dans la communauté MIR (Music Information Retrieval). Ces travaux se focalisent essentiellement sur la classification automatique des données et sur la visualisation des classes produites ou des proximités calculées (Li *et al.*, 2012). Mais, comme le souligne un article récent au titre explicite « The neglected user » (Schedl *et al.*, 2013), l'utilisateur est peu pris en compte dans la démarche. Il ne l'est souvent qu'*a posteriori* pour les tests de validation d'interfaces et son degré de liberté dans la gestion des données est fortement contraint par les topologies des espaces de classement définis dans les outils et les modalités d'interaction. De plus, comme le souligne Jones (Jones, 2008) il faut distinguer l'organisation de la gestion de l'information et la majorité des travaux porte plus sur la question de la gestion des données pilotée par des interfaces élaborées que sur celle de l'organisation personnelle sur des supports familiers.

En fait, à notre connaissance, la composition des discothèques numériques personnelles a été peu abordée dans la littérature (Sease & McDonald, 2009; Brinegar & Capra, 2010; Kamalzadeh *et al.*, 2012; Lee & Waterman, 2012; Jacques, 2015). Les données sur lesquelles ces travaux reposent sont issues d'enquêtes par questionnaires complétées souvent par quelques entretiens ethnographiques. Les observations confirment la difficulté d'estimation de la taille des collections à l'échelle individuelle et la variabilité des situations observées. Si ces enquêtes permettent de commencer à défricher les pratiques, elles se heurtent cependant aujourd'hui à deux écueils. La quantité d'informations auxquelles les natifs du numérique sont aujourd'hui quotidiennement exposés rend leur description verbale de plus en plus délicate. Toutes proportions gardées, on se retrouve face à un phénomène de type « big data » à l'échelle individuelle. De plus, avec les nouvelles modalités de « l'exposition de soi » sur les réseaux sociaux, il est légitime de s'interroger sur la véracité des contenus exprimés lors d'une enquête. Pour palier à ces difficultés, nous avons donc choisi ici d'observer *in silico* les contenus des disques durs des ordinateurs utilisés au quotidien sur lesquels on trouve des fichiers musicaux.

3 La méthodologie

Les données étudiées ont été recueillies auprès de 32 jeunes individus, pour la majorité étudiants, âgés de 17 à 30 ans, volontaires pour que le disque dur de leur ordinateur personnel principal (portable ou fixe) soit scanné pour le recueil des informations. Un outil logiciel parcourt un disque en enregistrant l'emplacement, le nom et la date de création de tous les fichiers musicaux pour les principaux formats et filtre les fichiers musicaux de petites tailles (< 1Mb) relatifs au système d'exploitation. Nous avons ainsi recueilli 10 4171 fichiers musicaux, soit une moyenne de 3255 morceaux par participant. Cependant, la moyenne est peu informative de par l'hétérogénéité de la distribution du nombre de fichiers/participants qui suit une loi proche d'une loi de Pareto.

Influencés par de nombreux travaux sur l'analyse de réseaux où la caractérisation de la topologie des relations tente de servir de révélateur à une organisation cachée plus subtile, nous avons commencé par étudier les arborescences des fichiers associés à chaque collection. Ces dernières se reconstruisent facilement à partir des données recueillies. Nous avons effectué une classification hiérarchique – non supervisée – de ces arborescences. Puis, nous avons tenté d'analyser le contenu – selon les genres musicaux – des différentes classes extraites de la classification et d'identifier les stratégies personnelles d'étiquetage des répertoires.

Analyse structurelle. Afin de découvrir les différentes stratégies mises en œuvre dans l'organisation des discothèques, nous avons effectué une classification des arbres à partir d'une description de ces derniers par des indicateurs combinatoires classiques : degré moyen des noeuds μ_d et écart-type associé, profondeur moyenne des nœuds μ_p et écart-type, longueur moyenne μ_l des chaînes entre la racine et les feuilles et écart-type, nombre de feuilles n_f (i.e nombre de fichiers musicaux) et nombre de nœuds n_r (i.e. nombre de répertoires). La dissimilarité entre deux collections est mesurée par la distance euclidienne entre les valeurs des indicateurs normalisés. Nous avions préalablement considéré une distance de Kullback-Leibler entre les distributions des différents indicateurs sans que cela n'apporte de précision supplémentaire dans les résultats obtenus. La classification a été effectuée par une méthode hiérarchique classique basée sur le critère de Ward.

Analyse "sémantique". Ayant récolté près de 100 000 noms de fichiers musicaux, une préclassification en un nombre de classes plus raisonnable nous est apparue indispensable pour une interprétation humaine. Un grand nombre des travaux de sociologie consacrés aux pratiques de l'écoute musicale dont nous avons eu connaissance font référence aux « goûts musicaux » et ces derniers sont définis par des « genres musicaux » . Dans les enquêtes par questionnaire classiques telles que celles menées par le Ministère de la Culture, une liste fermée de genres pré-définis est proposée. Mais, dans les fichiers numériques, le genre n'est pas donné ou est très mal renseigné et il nous a fallu recourir à une étape de classification. Malgré le développement de nombreux modèles computationnels, le problème de la détermination automatique du genre d'un morceau musical reste un problème délicat largement ouvert (Seyerlehner *et al.*, 2010), et à notre connaissance, les musicologues n'ont pas produit une ontologie consensuelle des genres musicaux qui pourrait servir de support à un codage automatique. Ce problème sortant largement du cadre de ce travail, nous avons choisi, pour des raisons opérationnelles, de recourir à la

base de données d'EchoNest [1] qui propose un étiquetage automatique des artistes en genres (sur une base de plus de 700 genres). L'extraction des noms d'artistes de nos données a été facilitée par les méta-données et nous avons ainsi recueilli 5405 noms d'artistes différents. L'étiquetage d'EchoNest repose principalement sur une analyse fréquentielle des termes provenant de publications en rapport avec les artistes collectées principalement sur des sites spécialisés et des réseaux sociaux. À chaque artiste apparaissant dans les métadonnées, nous avons associé via la consultation d'EchoNest le genre indiqué et nous avons conservé *in fine* pour chaque artiste le genre majoritairement associé sur l'ensemble des collections ; à chaque artiste est donc associé un seul genre musical. Sur notre échantillon, 478 genres musicaux sont représentés avec une moyenne de 67 genres par discothèque et une forte disparité.

Etiquetage manuel des répertoires dans les discothèques personnelles. En sus de la classification automatique des contenus en genres musicaux, nous avons tenté de mieux comprendre la procédure manuelle de classement en analysant les noms donnés aux répertoires. La très forte variabilité à la fois des noms proposés et de leurs orthographes a rendu la tâche difficile. En nous basant sur les résultats des enquêtes citées dans l'état de l'art, nous nous sommes focalisés sur trois classes d'étiquettes : album, artiste et genre. L'identification des albums a été effectuée en deux étapes : un filtrage manuel basé sur nos connaissances et un filtrage automatique basé sur un étiquetage automatique classiquement utilisé sur les sites de téléchargement « artiste-album-titre » et la position du répertoire dans la hiérarchie. Bien que des erreurs puissent persister, les différences entre les résultats sont suffisamment importantes pour ne pas altérer l'interprétation.

4 Classification des collections

La classification permet de faire apparaître trois classes distinctes correspondantes à des stratégies d'organisation des discothèques personnelles différentes :

— La classe 1, la plus peuplée, regroupe des discothèques basées sur un classement dichotomique qui combine des fichiers bien rangés et des gros répertoires « fourre-tout » dont le volume est supérieur à l'ensemble des autres catégories et contiennent deux fois plus d'artistes et de genres que tous les autres répertoires des discothèques de l'échantillon. De plus, dans ces discothèques les genres semblent plus spécialisés. Le nombre médian m_g de morceaux musicaux par genre dans toutes les discothèques est de 8 alors que dans la classe 1 seuls 24% des genres comportent plus de m_g morceaux. Un test de Wilcoxon confirme que cette différence est statistiquement significative.
— La classe 2 regroupe des discothèques de "power users" qui conservent un grand nombre de fichiers musicaux stockés dans une arborescence très structurée avec de très nombreux répertoires organisés eux-mêmes de façon récursive en sous-répertoires. Les discothèques de cette classe sont celles qui contiennent le plus grand nombre de fichiers musicaux (8408 en moyenne) et elles sont également les plus hétéroclites en terme de goûts puisque cette classe comporte 396 genres différents (40% de plus que la classe 1 et deux fois plus que la classe 3) et ces genres sont très fournis : 185 sont associés à plus de m_g fichiers avec une différence significative avec les autres classes.

1. ww.echonest.com

— La classe 3 se différencie des autres principalement par la faible profondeur de ses répertoires. Une analyse complémentaire des dates de création de ces derniers montre qu'en moyenne 56% de ces répertoires contiennent des fichiers ajoutés au même moment — contre 27% pour les autres classes—. Il s'agit donc d'un comportement de classement qui suit la stratégie « je range tout de suite en créant une nouvelle classe à chaque fois ». Et les discothèques de cette classe sont celles qui comportent le moins de genres différents. Une analyse des genres les plus représentés relève des genres qui sortent plus des courants dominants.

Cette typologie des comportements de classement a été complétée par l'analyse des noms donnés aux répertoires. L'album joue un rôle majeur dans le processus de classement ; ce qui peut différer des données verbales recueillies dans les enquêtes par questionnaire (Kamalzadeh *et al.*, 2012) où le genre a un rôle souvent beaucoup plus important. La typologie classique en trois items (genre, artiste, album) ne concerne que 80% des étiquettes et on note l'émergence de noms liés aux actions liés à la musique (« pour danser », « pour courir »). Cette pratique rappelle celle qui avait été déjà été mise en évidence avec des documents physiques où les attributs du classement ne dépendaient pas uniquement du contenu « objectif » des documents (e.g. auteur, sujet, titre) mais de l'interaction entre l'utilisateur et l'information (Kwasnik, 1991).

Cette première analyse permet de révéler des stratégies d'organisation différentes. Dans la littérature, qu'il s'agisse de classements d'artéfacts physiques ou numériques, bon nombre de travaux se réfèrent à la distinction introduite dans les travaux pionniers de Malone (Malone, 1983) sur l'organisation des documents au bureau entre les dossiers (« files »), où les documents sont étiquetés et ordonnés, et les piles (« piles ») où les documents sont posés sans ordre spécifique. Quel que soit le type d'information, la création de piles demande un effort minimal alors que la création de dossiers requiert un effort cognitif et manuel plus important (Jones, 2008). Nos résultats recouvrent la distinction classique (« neat » et « messy ») et l'affinent. Au niveau individuel une organisation méticuleuse peut se combiner avec une organisation négligée. Au niveau de l'échantillon, si des différences persistent, la combinatoire des modes d'organisation est finalement assez restreinte ; ce qui rend la perspective d'une assistance personnalisée plus aisée.

5 Conclusion

Dans cet article, nous avons proposé une analyse descriptive de l'organisation des collections musicales à partir de l'observation d'empreintes numériques (ici les arborescences de fichiers musicaux stockées sur les disques durs personnels). Il s'agit d'une première étape visant à identifier les processus mis en œuvre dans un environnement numérique familier. Nous avons mis en évidence différents modes de classement dont la robustesse devra être testée sur d'autres échantillons. Mais les facteurs explicatifs restent largement à explorer. La ré-utilisabilité reste une motivation souvent énoncée dans la littérature mais elle n'est pas la seule (Kaye *et al.*, 2006). L'identification des liens ou de l'absence de lien entre les modes de classement et les objectifs de l'utilisateur reste évidemment une question essentielle pour le développement d'une Ingénierie des Connaissances Personnelles.

Pour notre étude, nous nous sommes restreints aux données stockées sur un type de support spécifique mais aujourd'hui le nombre de supports qu'un individu a à sa disposition ne

cesse de croître, et s'ajoute aux dispositifs personnels, le stockage distribué sur des supports externes. Une communication récente de S. Abiteboul (détaillée dans Abiteboul *et al.* (2015)) au titre évocateur « Quand nos vies numériques deviennent des bases des connaissances » ouvre des nouveaux défis passionnants pour l'Ingénierie des Connaissances. La construction de ces ontologies personnelles revisite en profondeur la démarche classique dont la légitimité repose souvent sur un consensus entre experts d'un domaine. Et elle renouvelle l'intérêt d'une construction automatique qui reste encore aujourd'hui bien délicate.

Références

ABITEBOUL S., ANDRÉ B. & KAPLAN D. (2015). Managing your digital life. *Commun. ACM*, **58**(5), 32–35.

BRINEGAR J. & CAPRA R. (2010). Understanding personal digital music collections. *Proc. of the American Society for Information Science and Technology*, **47**(1), 1–2.

CHARLET J. (2005). L'ingénierie des connaissances, entre science de l'information et science de gestion. In *Entre la connaissance et l'organisation, l'activité collective*, p. 306–309. P. Lorino & R. Teulier Eds, La Découverte.

DENORA T. (2000). *Music in Everyday Life*. Cambridge University Press.

JACQUES J. (2015). Les pratiques d'organisation des collections musicales numériques comme enjeu central de l'écoute contemporaine. In *Colloque Musimorphose, Paris (actes à paraître)*.

JONES W. (2008). *Keeping Found Things Found : The Study and Practice of Personal Information Management : The Study and Practice of Personal Information Management*. San Francisco, CA, USA : Morgan Kaufmann Publishers Inc.

KAMALZADEH M., BAUR D. & MÖLLER T. (2012). A survey on music listening and management behaviours. In *Proc. of the 13th Int. Symp. on Music Information Retrieval*, p. 373–378.

KAYE J. J., VERTESI J., AVERY S., DAFOE A., DAVID S., ONAGA L., ROSERO I. & PINCH T. (2006). To have and to hold : Exploring the personal archive. In *Proc. of the SIGCHI Conference on Human Factors in Computing Systems*, p. 275–284 : ACM.

KWASNIK B. H. (1991). The importance of factors that are not document attributes in the organization of personal documents. *The School of Information Studies Faculty Scholarship*.

LEE C. (2011). *I, Digital : Personal Collections in the Digital Era*. The Society of American Archivist.

LEE J. H. & WATERMAN N. M. (2012). Understanding user requirements for music information services. In *Proc. of the 12th Int. Symp. on Music Information Retreival*, p. 253–258.

LI T., OGIHARA M. & TZANETAKIS G. (2012). *Music Data Mining*. Chapman & Hall/CRC.

MALONE T. W. (1983). How do people organize their desks ? implications for the design of office information systems. *ACM Trans. Inf. Syst.*, **1**(1), 99–112.

SCHEDL M., FLEXER A. & URBANO J. (2013). The neglected user in music information retrieval research. *Journal of Intelligent Information Systems*, **41**(3), 523–539.

SEASE R. & MCDONALD D. W. (2009). Musical fingerprints : collaboration around home media collections. In *Proc. of the ACM international conference on Supporting group work*, p. 331–340.

SEYERLEHNER K., WIDMER G. & KNEES P. (2010). A comparison of human, automatic and collaborative music genre classification and user centric evaluation of genre classification systems. In *Adaptive Multimedia Retrieval. Context, Exploration, and Fusion, 2010*, p. 118–131.